Michael Cohen
The University of Aizu
Tsuruga Ikki-machi
Aizu-Wakamatsu
Fukushima-ken 965, Japan

Integrating Graphic and Audio Windows

Abstract

It is important to exploit sound as a vital communication channel for computer–human interfaces. Developing this potential motivates both developing expressive models unique to audio and also exploring analogues to visual modes of representation. This paper elaborates an organization of presentation and control that implements a flexible sound management system called "audio windows." After reviewing audio imaging, spatial sound, and relevant underlying technology, an audio windowing prototype is described, implementing an extended model of free-field, two-dimensional spatial sound control. The system, "MAW" (acronymic for *m*ultidimensional *a*udio *w*indows), is a GUI (graphic *u*ser *i*nterface), integrating a graphic editor with a multidimensional spatial sound engine. Standard idioms for WIMP (*w*indow, *i*con, *m*enu, *p*ointing device) systems are reinterpreted for audio window applications, including provisions for directionalized and nonatomic spatial sound objects. Unique features include draggably rotating icons, clusters, dynamically collapsible hierarchical groups of spatial sound objects, and an autofocus mode that is used to disambiguate multiple presence.

1 Introduction

It is important to exploit sound as a vital communication channel for computer–human interfaces. Developing this potential motivates both developing expressive models unique to audio and also exploring analogues to visual modes of representation. Audio windowing is conceived of as a front end, or user interface, to an audio system with a spatial sound backend. This applications note surveys the ideas underlying audio windowing and describes a design for a system investigating these ideas. The paper addresses these questions: What features of a GUI (*g*raphic *u*ser *i*nterface) can be exploited by an audio window system? How can an audio window system be built on a spatial sound system? How can gestural intuitions and user interface conventions be gracefully adapted and extended to support audio windows? After tracing some underlying technology of audio imaging in computer–human interfaces, these questions are answered by the description of an audio windowing prototype, "MAW" (acronymic for *m*ultidimensional *a*udio *w*indows), an exocentric graphic mouse-based interface based on an extended model of free-field, two-dimensional (2D) spatial sound.

1.1 Audio Imaging

Part of listening to an admixture of conversation or music is being able to appreciate the overall blend while also being able to hear the individual voices

Presence, *Volume 1, Number 4, Fall 1992*
© *1993 The Massachusetts Institute of Technology*

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 001

or instruments separately. This synthesis/decomposition duality has the opposite effect of masking: instead of sounds hiding each other, they are complementary and individually perceivable. In music, for instance, instruments of contrasting color are used against each other. Localization effects contribute to this antimasking by helping the listener distinguish separate sources, be they instruments in an ensemble or voices in the cacophony of a cocktail party (Koenig, 1950; Blauert, 1983).

Audio imaging is the creation of sonic illusions by manipulation of separate channels. For instance, when music (a good recording of classical music, say) is played, the music from different instruments comes from distinctly different directions. The violins seem to be on the listener's left, the cellos and double basses are on the right, the violas face the listener, and the percussion, woodwinds, and brass are to the rear of the orchestra. On a stereo system, the sound really comes from only the left and right transducers, whether headphones or loudspeakers. Current audio systems project only a one-dimensional arrangement of the real or mixed sources. In traditional sound reproduction, the apparent direction from which a sound emanates is typically controlled by shifting the balance of the unmodified sound source between left and right channels. But this technique yields images that are diffuse, and located only between the speakers.

1.2 Spatial Sound

Given multiple audio channels, a display system needs a way of perceptually segmenting them, a way of distinguishing them from each other or segregating the streams. A simple way is by just making the channels of interest louder than their siblings. A more sophisticated approach, spatial sound, employs technology that allows sound sources to have not only a left–right attribute (as in a conventional stereo mix), but up–down and back–forth qualities as well.

Spatial sound encourages auditory localization, and separately spatialized channels invoke the "cocktail party effect" (Cherry, 1953). In the absence of visual cues, someone listening to music distinguishes the channels (voices, instruments, parts) by position, tone/timbre, and melodic line and rhythm. Similarly, at a party with many simultaneous conversations, a mingler can still follow any particular exchange by filtering according to position, speaker voice, and subject matter. The last two dimensions are not controlled by spatial sound systems, but by adjusting the virtual position of the sources with respect to the listener, other aspects of perceptual segmentation are implicitly enabled. Spatial sound projects audio media into space by manipulating sound sources so that they assume virtual positions, mapping them into a perceptual envelope around the listener. These virtual positions allow a psychological separation of the channels, via space-domain multiplexing.

By creating psychoacoustic effects (usually with DSP, *d*igital *s*ignal *p*rocessing), many scientists are developing ways of generating and controlling this multidimensional sound imagery (Chowning, 1977; Moore, 1983; Fisher, McGreevy, Humpries, & Robinett, 1986; Kendall, Martens, Freed, Ludwig, & Karstens, 1986a,b; Martel, 1986; Martens, 1987; Myers Laboratories, 1987; Kendall & Freed, 1988; Wenzel, Wightman, & Foster, 1988a,b; Bernardini & Otto, 1989; Scott, 1989; Sorkin, Wightman, Kistler, & Elvers, 1989; Begault & Wenzel, 1992; Fisher, 1990; Kendall, Martens, & Wilde, 1990; Loomis, Hebert, & Cininelli, 1990; Setton, 1990; Wenzel, Stone, Fisher, & Foster, 1990; Wenzel, Wightman, & Kistler, 1991). The goal of a sound spatializer is to create the impression that the sound is coming from different sources and different places, just like one would hear "in person." Spatial hearing can be stimulated by assigning each source a virtual position with respect to the listener and simulating the auditory positional cues. A display based on this technology exploits the human ability to quickly and subconsciously localize sound sources.

Binaural (stereo) localization cues include *i*nteraural *t*ime (phase) *d*elay (ITD), a function of the interaural distance; *i*nteraural *i*ntensity *d*ifference (IID), a consequence of the acoustic head shadow; and notches in the binaural spectral response (Blauert, 1983; Martens, 1989). One approach to spatial sound uses a (hardware- or software-based) convolution engine, which incorporates *f*inite *i*mpulse *r*esponse (FIR, also known as tapped delay) digital audio filters, whose output signals are presented to

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 002

stereo loudspeakers or headphones. Binaural localization cues may be captured by *h*ead-*r*elated *t*ransfer *f*unctions (HRTFs), impulse responses measured for the head and pinna (outer ear) of human or dummy heads. For each spherical direction, a left–right pair of these transfer functions is measured and stored as FIR filter coefficients. Sound can then be spatialized by driving digitized input signals through these HRTFs in a convolution engine, creating psychoacoustic localization effects by introducing binaural spatial cues into an originally monaural channel, "placing" the channel within the perceptual space of the user (Kendall & Martens, 1984; Gehring, 1987; McKinley & Ericson, 1988; Richter & Persterer, 1989; Wenzel, 1992).

1.3 Audio Windows

"Audio windows" is an auditory-object manager, a user interface, or frontend, to a spatial sound system. The generalized control model of a window here is by analogy to graphic windows, as in a desktop metaphor: an organizational vehicle in the interface that does not directly involve room acoustics. Researchers have been studying applications and implementation techniques of audio windows for use in providing multimedia communications (Ludwig & Pincever, 1989; Ludwig et al., 1990; Cohen & Ludwig, 1991a,b; Cohen & Koizumi, 1991b). The general idea is to permit multiple simultaneous audio sources,[1] such as in a teleconference, to coexist in a modifiable display without clutter or user stress. The distribution of sounds in space is intended to realize some of the same kinds of benefits achieved by distribution of visual objects in graphic user interfaces.

A powerful user interface would allow arbitrarily setting and adjusting the position of the audio channels, so the positions of the listener and sound sources may be constantly changing as they move around each other and a virtual room. By using an audio window system as a binaural directional mixing console, a multidimensional pan pot (a *pan*oramic *pot*entiometer controls the placement of a channel in a conventional stereo mix), users could set parameters reflecting these positions. Listeners altering these parameters could experience the sensation of wandering around a conference room, among the teleconferees. Members of the audience at a live or recorded concert could actively focus on a particular channel by sonically hovering over the shoulder of a musician in a virtual concert hall. Minglers at a virtual cocktail party might freely circulate. Sound presented in this dynamically spatial fashion is as different from conventional mixes as sculpture is from painting.

1. Since the word "speaker" is overloaded, meaning both "loudspeaker" and "talker," this paper uses "source" to mean a logical sound emitter. Similarly, "sink" is used to describe a virtual listener, a logical sound receiver.

2 MAW: Multidimensional Audio Windows

The standard pointing device in a *w*indow/*i*con/*m*enu/*p*ointing device (WIMP) system is a mouse. The desktop metaphor and conventions for WIMP interfaces were pioneered by the Xerox Star (Johnson, Roberts, Verplank, Smith, Irby, Beard, & Mackey, 1989), but the "look-and-feel" was emulated and extended by systems that include Apple's Lisa and Macintosh, MIT's X-Windows, Sun's Views, Tools, and NeWS, NeXT's NextStep, and Microsoft's Windows. A mouse, trackball (a mouse on its back), or joystick would suffice for positioning objects in 2-space. One can imagine direct control of position (with a mouse, trackball, or DataGlove), steering control (with a return-to-center joystick), or programmed control (using automatic circulation).

As long as visual tokens for the spatial sound objects are accessible through the workstation, they can be manipulated through the operations described above, via a suite of direct and indirect manipulation techniques (Cohen & Koizumi, 1992a). For example, a mouse-based system realizes indicating, highlighting, and relocating as pointing, clicking, and dragging. This paper describes a system that implements graphic control of audio windows, by reinterpreting WIMP conventions to handle an extended spatial sound model.

2.1 Implementation

MAW is an interactive audio windowing and 2D spatial sound system, suitable as a teleconferencing sys-

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 003

tem, as well as for asynchronous applications like voicemail. Rather than relying on disembodied sources, MAW includes a GUI, employing visual representations of sound elements manipulated through a workstation, integrating graphic and audio windows by manipulating spatial audio channels via graphic control.

MAW is implemented on a NeXT computer, under the NextStep environment. The program consists of about 20,000 lines of Objective-C and Postscript code, plus associated graphics and graphic configurations. MAW's interface reinterprets the standard WIMP idioms, which are derived from a long ancestry of graphic programs. MAW is platformed upon Draw, an object-oriented drawing editor by Paul Hegarty, from which it inherits much of its graphic functionality and code. (As much as possible, the description that follows outlines only the unique features of MAW, the enhancements that make it appropriate for audio windowing.)

MAW's native (self-contained) backend can spatialize a single channel. The Focal Point™ (Gehring, 1988, 1990) multidimensional audio software module for the NeXT, designed by Bo Gehring, is adapted from the Focal Point audio display software for the Macintosh II. As described earlier, the underlying principle of operation is that the input audio sound stream is convolved with transfer functions corresponding to the desired perceived source direction.

As shown in Figure 1, an audio stream is sampled by a Metaresearch Digital Ears module (Metaresearch, 1991), a high-speed, high-precision A/D converter. The convolution engine is the NeXT's Motorola 56001 DSP, programmed to filter the mono input signal into directionalized stereo signals, by convolving it with frontend-selected pairs of transfer functions, which emulate relative positional cues. These HRTFs are adapted from dummy-head recordings[2] of impulse responses. The binaural convolution is carried out simultaneously for the left and right ears; the result is a stereo pair which can be heard through headphones (or with less accuracy through loudspeakers). Focal Point's standard sampling resolution is 16 bits, at a sampling rate of 44.1 kHz, with tap lengths of about 100 samples.

Upon invocation of MAW, Focal Point software is downloaded into the NeXT DSP and then updated only when a source or sink is moved; audio processing continues without host CPU interaction. Smooth transitions are provided between relative locations distributed every 10° in the horizontal plane. Updating the transfer function in the DSP has a latency of about 3 msec, which compares favorably with known visual displays. This update rate (which can be in excess of 300 Hz) is well suited to rapid source motion with respect to the listener (as in other applications where a headtracking feature is included).

MAW also features two techniques for spatializing multiple channels: one is the ability to locally drive multiple alternate external spatializers, assigning logical channels to heterogeneous physical devices via a preferences (control) panel. MAW currently has drivers for its internal Focal Point and the external Stork and SIM**2 (acronymic for *s*ound *im*age *sim*ulator) convolution engines, as well as hooks for Macintosh-based Focal Point and the Convolvotron (two commercially available spatialization engines). Further, an arbitrary number of NeXTs, each running MAW coupled under an extension of Greyboard (a groupware utility written by Joe Freeman), may convene in single or multiple conferences, each workstation using its own spatializer. The outputs of each workstation can then be mixed to form multiple, spatial audio channels, as shown in Figure 2.

2.2 Virtual Room View

The graphic representation of MAW's virtual room is an aerial projection, a 2D bird's-eye view. This perspective flattening was implemented partly because of its suitability for visual display on a workstation monitor. Figure 3 shows a typical view of such an overhead representation, along with the border, buttons, and scrollers that make it a NextStep window, as part of a snapshot of a typical session. Such a planar slice was also chosen to maximize positional accuracy for MAW users; localization cues for virtual sound sources are more robust from person to person for the azimuthal dimension than for

2. MAW's version of Focal Point (which is different from Focal Point on a Macintosh) uses measurements made by the Computer Music Group at Northwestern University with a (KEMAR) mannequin.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 004

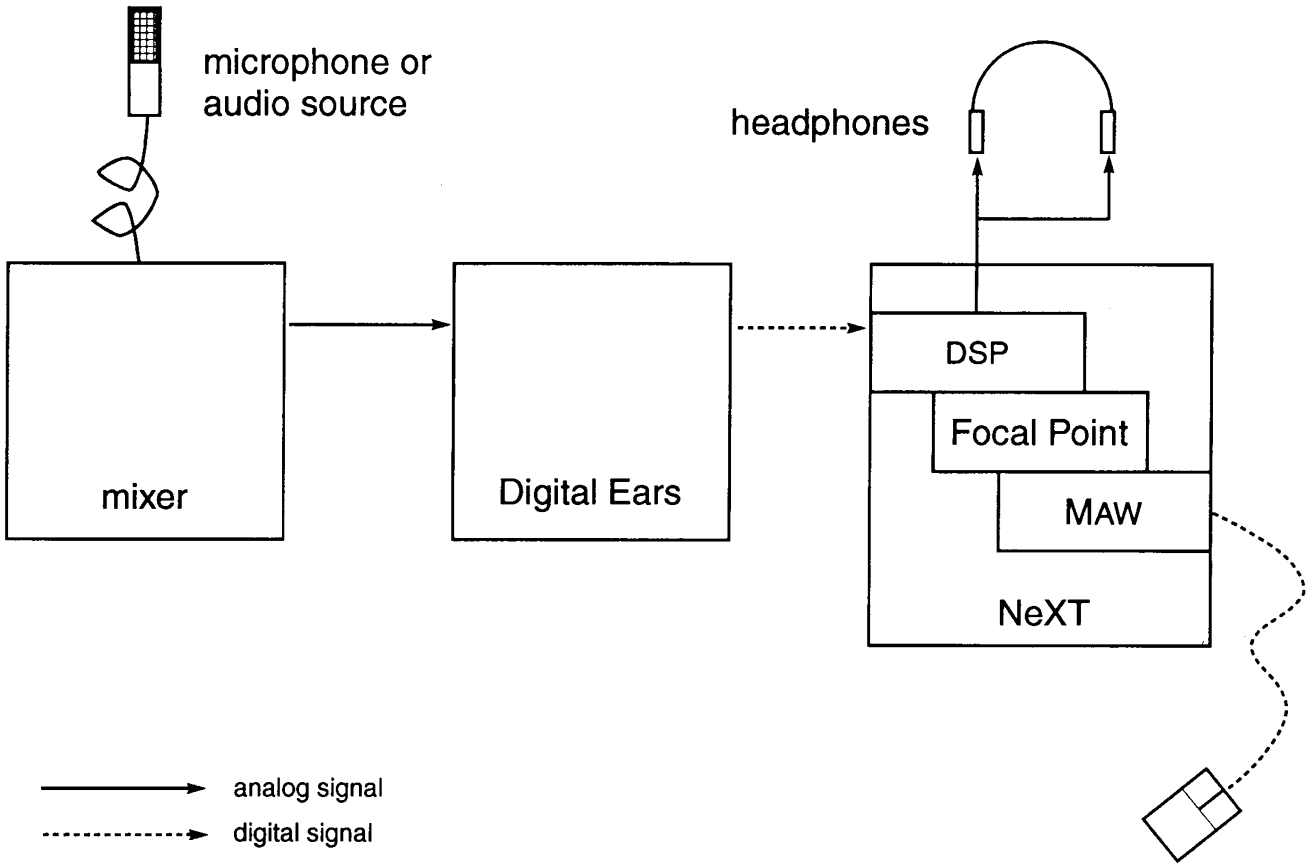

Figure 1. *Self-contained system architecture.*

the elevational (Wenzel et al., 1988b, 1991). Azimuthal cues are based primarily on interaural differences due to the horizontal arrangement of two spatially separated ears, while elevational cues are based on the complex spectral shapes resulting from the interaction of incoming sound waves with each listener's pinna structures. Since outer ear shapes presumably vary more from person to person than head size, it is not surprising that individual differences in the ability to localize virtual sources tends to be greatest in elevation, especially when the synthesis is based on nonindividualized HRTFs (Wenzel et al., 1991).

2.3 Rotating Icon Representation

Sources and sinks may correspond not only to people, but also to inanimate objects or clusters of objects, and no attempt is made to distinguish among the different manifestations; MAW adopts the simplification that all spatial sound objects are at once 2D sources and potential sinks, arrayed in the horizontal plane. The system allows a variety of iconic styles, including any combination of static pictures [which might be head-and-shoulder snapshots (as in Fig. 3) or top-down (as in Fig. 6)], symbolic icons (as in Fig. 6), overlaying sound field density plots (as in the right column of Fig. 4), and textual labels (as in Fig. 3). These representations may be switched "on-the fly," in response to user preference.[3] Spatial sound objects in MAW have both rectangular co-

3. For example, text labels might be blanked, or presented in alternate languages. Head-and-shoulders snapshots allow easy recognition, but top-down views are more consistent with the room's perspective.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 005

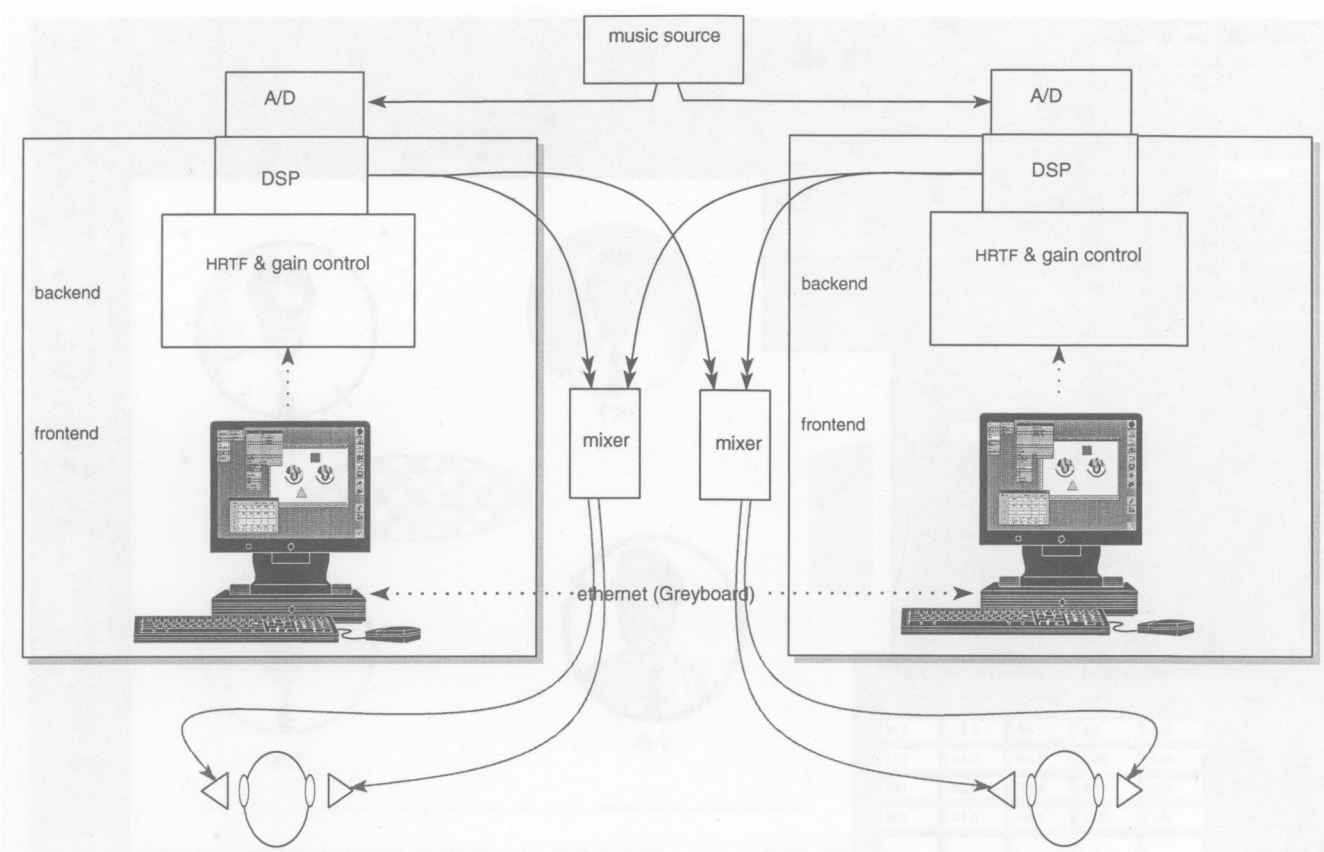

Figure 2. *Double channel spatialization (without external convolution).*

ordinates and angular attributes. Icons for these nonomnidirectional sources and sinks indicate both their mutual positions in virtual space and their orientations.

Sinks and sources may be repositioned, individually or in groups. (Association is discussed later in Section 2.5.) Since the spatial sound object states include both angular and rectangular attributes, they may be rotated as well as translated. This repositioning may be by either direct or indirect manipulation, may be (rectilinearly or latticed) constrained or unconstrained, may be continuous or discrete, nonconvergent or convergent to some target, and may be absolute or relative to the starting position of the icon (Cohen & Koizumi, 1992a).

The most direct style of relocation is by repositioning (singly or in groups) the icons with the mouse. For ordinary graphical objects, the relocation cursor is a flat hand; for spatial sound icons, the relocation cursor becomes a closed hand with index finger extended. This cursor suggests pivotability around the "hot spot" at the tip of the finger. Instead of simply translating rectilinearly, MAW's spatial sound icons rotate as well, simulating a natural object dragged across a plane. Rather than using a trackball or 3D joystick to control planar orientation of the objects (x–y position specified normally, and azimuth specified with the torsional z), the icons swing around in the direction of mouse dragging, depending on where they are grabbed, in an intuitive repositioning that integrates translation and rotation.

2.4 Gain Control

A similar suite of icon manipulation techniques can be used to adjust the gain of the channels. The user can resize a selected atomic spatial sound icon by dragging one of the resize handles (knobs) on its bounding rectangle (as in the bottom left icon of Fig. 3), invoking a

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 006

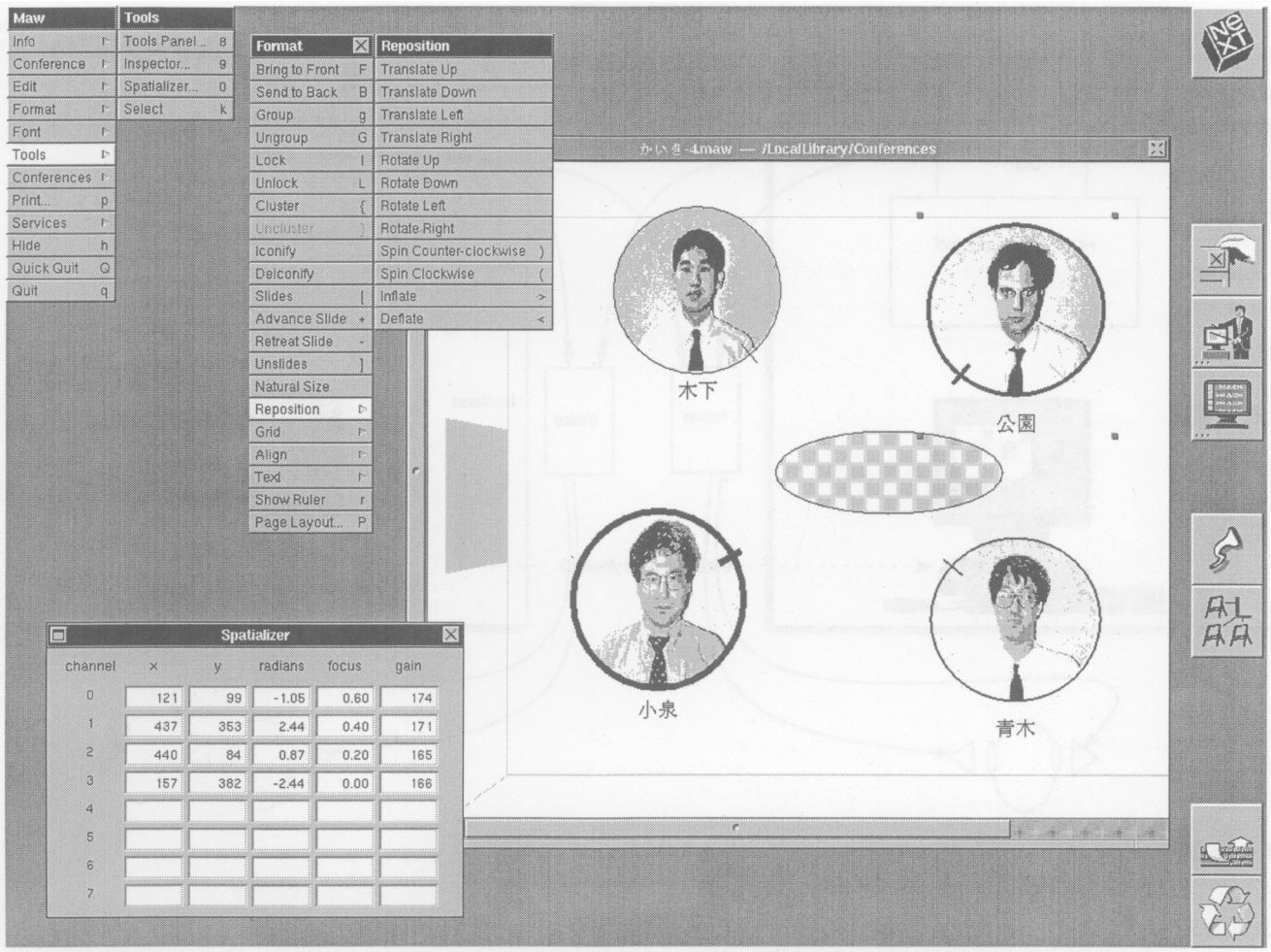

Figure 3. *Screen shot.*

menu or command key operation, or by editing a numeric cell corresponding to its size in a panel. The size of a source corresponds to individual gain; the size of a sink corresponds to general gain. For the sake of parsimony, iconic size (along one linear[4] dimension) is used as a determinant of both metaphorical ear and mouth size. Gain is proportional to the size of the source's mouth (amplification) and the sink's ear (sensitivity), so enlarging an icon makes its owner both louder and more acute. Thus, to make a *single channel* louder or softer, a user simply resizes the respective icon, but to make *everyone* louder or softer, the user need only resize his/her own icon. Gain is also proportional to the inverse square of the distance between the sink and source, so another way to change perceived volume is to have the source and sink approach or recede from each other.

Spatial sound sinks have an orientation established by their positions relative to the sources. Furthermore, sources are also directed, allowing, say, human speakers to project their voices, addressing a particular sector of an audience. With a directional radiation pattern, a source's beaming can be attenuated positionally. Sound may be projected narrowly (a focused whisper), multicast to a target group, or broadcast to some universe

4. The linear dimension that actually determines the gain is (arbitrarily) the width, but since icons are almost always resized while holding the aspect ratio fixed (to avoid graphic distortion), height (or diagonal, or circumference . . .) would have worked equally well.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 007

(conference, concert, company, country, etc.). Beyond the fact that nonomnidirectional virtual sources better approximate "real" sources, the goal is less to mimic natural reality than to expand the range of expressive options.

A modified frequency-independent cardioidal pattern is used to model the sound field radiation of the nonomnidirectional sources. The chosen relationship specifies an azimuth-dependent beaming of the speaker, an idealized directional pattern, with exaggerated distance effects. The overall amplitude of a source → sink transmission is independent of the sink's transfer function, and is determined (Cohen & Koizumi, 1991a,b) by

$$gain = \begin{cases} \dfrac{1 + (focus_{\text{source}})[\cos(\theta)]}{1 + focus_{\text{source}}} \cdot \dfrac{(size_{\text{source}})(size_{\text{sink}})}{distance^2} & \text{far-field} \\ \dfrac{1 + (focus_{\text{source}})[\cos(\theta)]}{1 + focus_{\text{source}}} & \text{near-field} \end{cases} \quad (1)$$

where $0 \leq focus < 1$, θ is the rotated angle of inclination of the sink with respect to the source, and *distance* is between the source and the sink, in the same units as the respective *sizes*.

The *focus* represents the dispersion, or beaming, of the sound. The first term of the extended Eq. (1) captures the rotational effects of the source; the normalized fraction uses the *focus* coefficient to scale the contribution of the azimuth. For a *focus* of zero, the radiation pattern is omnidirectional (as the first terms of the extended expression reduce to unity). A *focus* of greater than zero enables a cardioidal pattern, as if the source were using a megaphone. As two objects approach each other, the second far-field fraction factor approaches unity, and is rectified to unity as the objects osculate and overlap, invoking near-field behavior. The icon shown in the top left of Figure 4 has *focus* = 0.1, and a corresponding radiation pattern (sound field density, in which lighter areas indicate greater intensity) in the top right is almost omnidirectional. In contrast, the icon in the bottom left has *focus* = 0.9 (as indicated by its thicker arm), and its radiation pattern in the bottom right is almost perfectly quiet in the shadow behind the head.

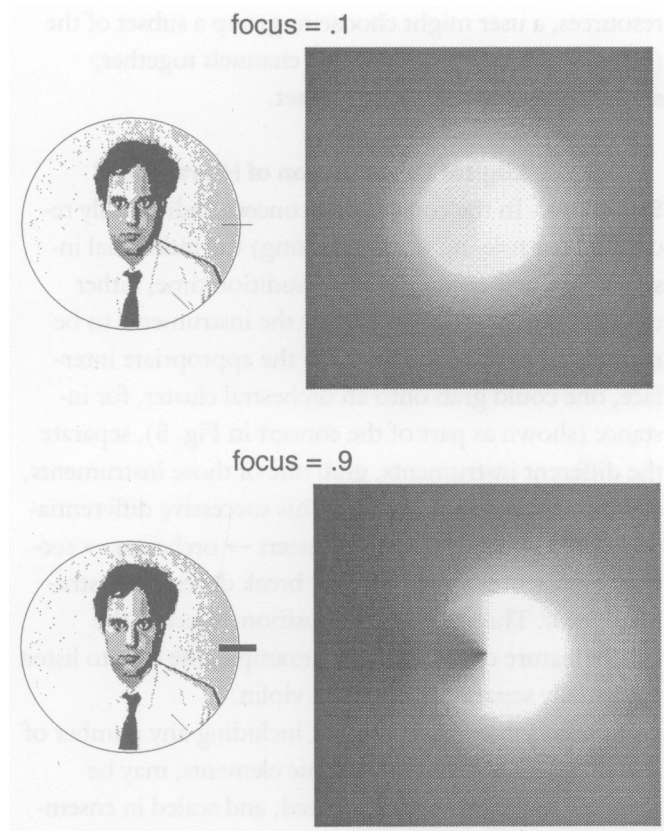

Figure 4. *Icons and radiation patterns.*

2.5 Clustering and Unclustering

An interface with the ability to model sources hierarchically would allow decomposition into their atomic components. The motivation for such a cluster feature is twofold.

2.5.1 Conservation of Spatializer Resources. Postulating a switching matrix on either side of the spatializer processor (Cohen & Ludwig, 1991a,b), along with dynamic allocation of spatializer channels, a cluster feature organizes separate input channels that share a single spatializing channel. One application might involve zooming effects. Distant sources would not be displayed; but as it approaches, a cluster would appear as a single point; only to disassociate and distribute spatially as it gets closer. This focus allows navigation in arbitrarily large spaces, assuming a limited density of point sources. Alternatively, with limited spatializing

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 008

resources, a user might choose to group a subset of the (less pleasant or less important) channels together, stacking them in a corner or closet.

2.5.2 Logical Organization of Hierarchical Structure. In the context of a concert, individually recording (or mic-ing or synthesizing) the individual instruments, and mixing them at audition time, rather than in "postproduction," allow the instruments to be rearranged by the listener. With the appropriate interface, one could grab onto an orchestral cluster, for instance (shown as part of the concert in Fig. 5), separate the different instruments, grab one of those instruments, and move it across the room. This successive differentiation could go right through concert → orchestra → section → instrument and actually break down the instrument itself. This superdecomposition aspect of the cluster feature could allow, for example, the user to listen to spatially separate strings of a violin.

A collection of MAW objects, including any number of spatial sound objects and graphic elements, may be grouped and then moved, rotated, and scaled in ensemble (like a chorus of distinct but coupled voices). As a special case of grouping, MAW also implements clusters. When clustered, the selected spatial sound elements merge into one another, aligning and converging until they are spatially and graphically indistinguishable. Unlike members of a group that maintain their constellation, members of a cluster temporarily collapse into a bundle, subsuming their individual identities in a collective display. The last-selected icon is designated as the model: when the cluster operation is invoked, the other selected icons dynamically approach it, converging on its position, orientation, and display attributes.

Uncluster is the inverse of Cluster. When invoked, the penultimate level of nested components replaces their selected parent(s). Since clustered objects retain a memory of their former characteristics, including relative displacement from their (possibly relocated) parent cluster icon, unclustering restores these attributes, in the inverse (dynamically animated) operation. The animation comprises a (spatiotemporal) interpolation between the cluster and original states, restoring the objects to their natural attributes.

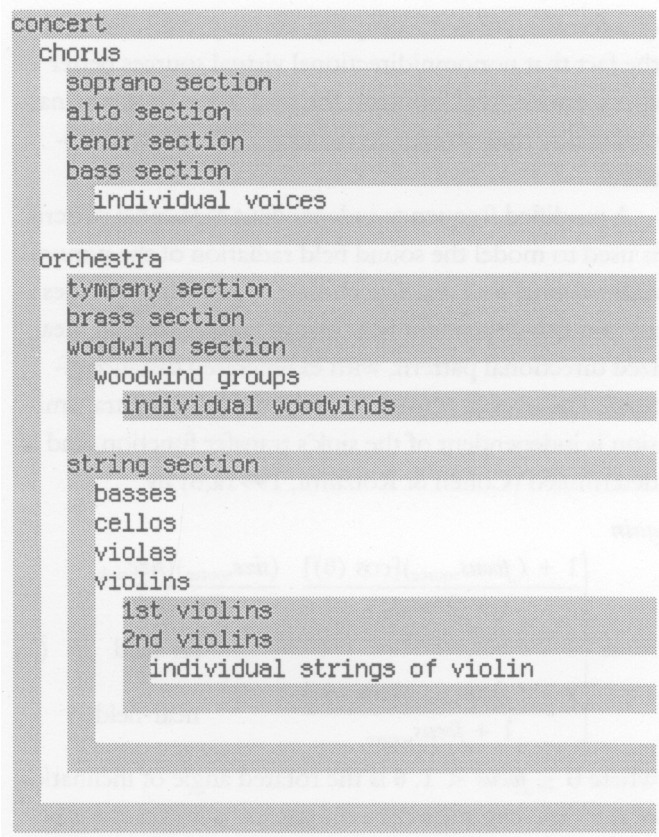

Figure 5. *Example of hierarchical decomposition of a concert.*

Unclustering can be likened to viewing the objects through a generalized fish eye lens (Furnas, 1986), which spatially warps the perception of the localized objects to enlarge an area of focus and shrink everything else. That is, when the user indicates a direction of special interest, the objects in that direction effectively approach the user and recede from each other in perspective. While the other objects do not get pushed into the background, the idea is the same: to effect an external rearrangement of sources that complements an internal reordering.

2.6 Schizophrenia and Autofocus

Some systems support multiple visual windows, each featuring a different perspective on a scene. In flight simulators, for example, these might be used to display (egocentric) views out cockpit windows, and/or views from a completely different location—high above the

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 009

airplane, for example, looking down (exocentrically): a virtual "out-of-body" experience. Since audition is (biasedly) omnidirectional, perhaps audio windows can be thought of as implicitly providing this multiperspective capability, audio sources being inherently superimposed. MAW also features a "schizophrenic" mode, allowing multiple sinks in the same or different conference rooms, explicitly overlaying multiple audio displays.

A simple teleconferencing configuration typically consists of several icons, representing the distributed users, moving around a shared conference space. Each of these icons represents a source, the voice of the associated user, as well as a sink, that user's ears. However, MAW allows users to have multiple sinks designated (through a preferences panel), effectively increasing their attendance in the conference. Such a feature might be used to pay close attention to multiple sources, even if those sources are not repositionable; just as in ordinary settings, social conventions might inhibit dragging someone else around a shared space. One could pay close attention to multiple instruments in a concert without rearranging the ensemble. One could leave a pair of ears in one conference, while sending another pair to a side caucus, even if the side caucus happens to be in the same room.

Since simultaneous and identical multiple icons introduce too many contradictions, icons are distinguished by unique channel numbers. Multiple sinks may be constructed a priori, or spawned ad hoc, perhaps by duplicating an original sink (then changing the channel number of the clone).

The apparent paradoxes of one's being in multiple places simultaneously are resolved by partitioning the sources across the sinks. If the sinks are distributed in separate conference rooms, each source is localized only with respect to the sink in the same room. If multiple sinks share a single conference room, an "autofocus" mode is employed by anticipating level difference localization, the tendency to perceive multiple identical sources in different locations as a single fused source. (This is related to the precedence, or Haas, effect.) Rather than adding or averaging the contribution of each source to the multiple sinks, MAW localizes each source only with respect to the best (loudest, as a function of distance and mutual gain, including focus and orientation) sink.

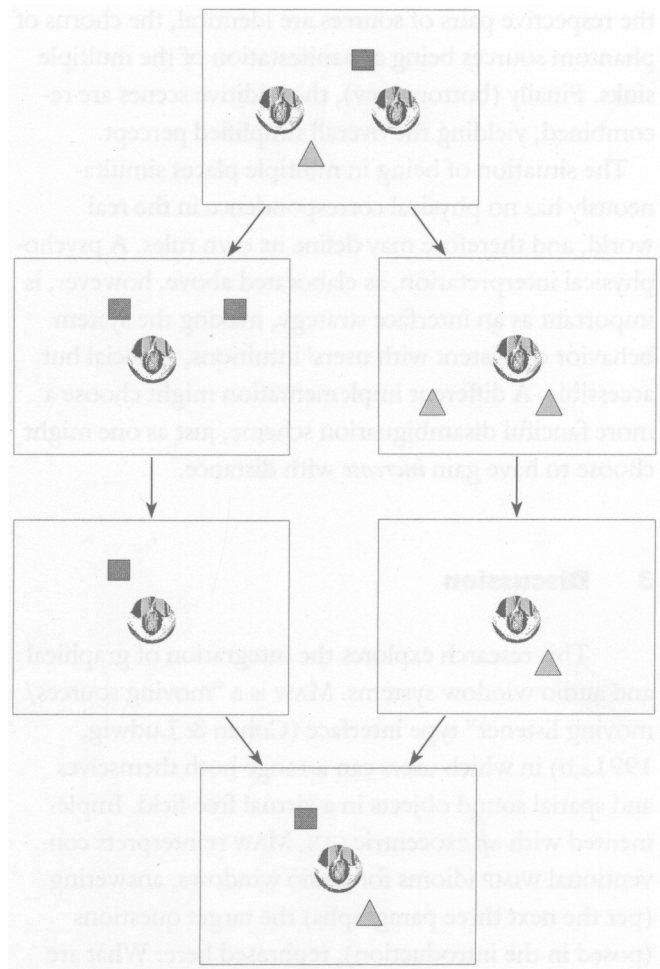

Figure 6. *Schizophrenic mode with autofocus.*

Figure 6 illustrates this behavior for a conference (top row) with two sinks, represented by top-down icons, and two different sources, represented by a square and a triangle. In the absence of room acoustics, multiple sinks perceiving a single source is equivalent, via "reciprocity" or symmetry, to a single sink perceiving multiple identical sources. Therefore the exampled scene can be decomposed source-wise into two additive scenes (second row), each single sink combining the parent sinks' perceptions of the respective sources. These configurations reduce (third row), via the "autofocus" level difference anticipation, to the respective sinks and only the loudest source. The loudest source is actually the closest, since

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 010

the respective pairs of sources are identical, the chorus of phantom sources being a manifestation of the multiple sinks. Finally (bottom row), the additive scenes are recombined, yielding the overall simplified percept.

The situation of being in multiple places simultaneously has no physical correspondence in the real world, and therefore may define its own rules. A psychophysical interpretation, as elaborated above, however, is important as an interface strategy, making the system behavior consistent with users' intuitions, artificial but accessible. A different implementation might choose a more fanciful disambiguation scheme, just as one might choose to have gain *increase* with distance.

3 Discussion

This research explores the integration of graphical and audio window systems. MAW is a "moving sources/moving listener" type interface (Cohen & Ludwig, 1991a,b) in which users can arrange both themselves and spatial sound objects in a virtual free-field. Implemented with an exocentric GUI, MAW reinterprets conventional WIMP idioms for audio windows, answering (per the next three paragraphs) the target questions (posed in the introduction), rephrased here: What are appropriate graphic paradigms for audio windowing systems? What would be useful features of an audio window backend? What extensions to current user interface idioms are needed to support audio windowing?

Spatial sound driven by a graphic exocentric display seems to be an effective implementation of an audio windowing system. MAW blurs the self/other distinction by iconifying all the users with similar tokens. A mouse- and monitor-driven interface allows manipulation of all the system entities; the metaphorical universe is projected onto an external and egalitarian medium. Menu commands, Command-keys, and extended arrow-key combinations span relative modes of relocation and reorientation. The dual of a reconfigurable graphic map and editable numeric panels span absolute modes of selection, resizing, relocation, and reorientation. MAW demonstrates the potential of audio windows as a way of organizing acoustic space, and the reinterpretation of WIMP idioms as an effective metaphor for such a system. It is designed to exploit innate localization abilities, our perception of spatial attributes, and our intuitive notions of how to select and manipulate objects distributed in 2-space. Everything in the system is manifest, and objects and actions can be understood purely in terms of their effect on the respective displays. Precise auditory localization is difficult, but informal experiments indicate that MAW's visual and acoustic displays compliment each other, a glance at the map disambiguating auditory cues.

MAW provides ways of manipulating the spatial sound objects as atoms, groups, and hierarchical trees (nested clusters). The cluster feature makes the interface not quite WYSIWYG (*w*hat *y*ou *s*ee *i*s *w*hat *y*ou *g*et), but this hiding of information is necessary for hierarchical abstraction. In order to support full-duplex conferencing, a large[5] number of audio channels must be channeled through audio imaging processors. Other applications, including voicemail, hypermedia, and music, require an arbitrarily large number of separately spatialized sonic channels. For all of these, and any task involving terminal audio management, spatial data organization, or scientific auralization (Kendall, 1990), a clustering mechanism is useful, both as a way of imposing a logical hierarchy on many sources, and, in conjunction with an audio switching matrix, as a way of conserving channels.

WIMP interface conventions can be seamlessly extended to manage 2D spatial sound objects, with the enhancement of draggably rotating icons, allowing non-omnidirectional sources and sinks; extension of chorded arrow combinations, enabling indirect spatial sound object manipulation (by allowing rotation, for example); and clusters, whose order-sensitive selection defines hierarchically collapsing groups of spatial sound objects. The direct manipulation object–command (noun–verb) syntax is well suited for moving sources and sinks. Mode-indicating cursors, animated icons (Baecker, Small & Mander, 1991) during dynamic tracking, and clustering effects contribute to the complete suite of manipulation techniques available.

5. Since every user must spatialize every other user's voice, the number of channels to spatialize grows quadratically, or as $O(|users|^2)$.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 011

MAW is intended to interpolate between conventional telephony and VR, but cannot be said to do more than suggest actual acoustic environments. Its major deficiency is the use of only two spatial dimensions, which seems to be more of a liability for musical expression than for teleconferences (as people seated in real conferences tend not to exploit the vertical dimension anyway). Further, simulating distance cues is a difficult and not-yet-solved problem that goes beyond MAW's simple gain changes. Besides peoples' natural inability to estimate distance with precision and MAW's deliberate exaggeration of distance effects, an inverse relation does not perfectly capture real effects (Blauert, 1983; Begault, 1991; Wenzel, 1992). MAW's modeling of source directionality is also not veridical: the selection of a cardioid is somewhat arbitrary, and a flat (frequency-independent) attenuation of gain is not the best model of a rotating source, which should change timbre as well as loudness. It would be better to have a second set of transfer functions that capture these shadow effects, and convolve the digitized source thrice: once for source rotation and twice (left and right ears) for sink revolution. MAW further oversimplifies reality by neglecting occlusion, the obstruction of a source's sound by other objects in the virtual room: Doppler shifts, the pitch bending exhibited by moving sources; indirect reflections (discrete echoes), the ratio of whose energy to that of direct sounds is another cue for estimating distance; and reverberation, statistically averaged room ambience. The absence of some of these cues is sometimes associated with perceptual errors like front ↔ back reversals (Foster, Wenzel & Taylor, 1991).

The explored models give users a telepresence into metaphorical space. Some of the features implemented find close approximation in visual or physical media; others elude analogy. But since it depends on perceptual rather than symbolic processing, MAW requires minimal instructions or training. It provides a testbed for exploring the potential of the emerging technology and for researching relevant human factors issues. Using systems like MAW for audio windowing can be likened to sonic (analytic) cubism: it presents several audio perspectives (on an assembled conference or concert) simultaneously.

As sound technology matures, and more and more audio and multimedia messages and sessions are sent and logged, the testimony of sound may come to rival that of the written word. Audio windows are a way of organizing and controlling sound. New media spend their early years recapitulating the modes of older media (McLuhan & Fiore, 1967); the project described by this paper hopes to abbreviate this phase for audio windows by accelerating its conceptual development. The issues involved in this research are both timeless and timely—timeless because the research analyzes the way people communicate with and through sound and space, and timely because the requisite technology has only recently ripened sufficiently for scientists to start exploring the expressive potential of the medium.

Acknowledgments

Paul Hegarty wrote Draw, the graphic editor on which MAW was platformed. Joe Freeman wrote Greyboard, MAW's conferencing utility. Bo Gehring designed Focal Point, the native DSP backend. Elizabeth Wenzel, Toni Cohen, and anonymous referees provided fertile comments. Some of the conceptual motivation for the research was developed at Bellcore under Chris Riley (Cohen, 1987) and Lester Ludwig (Cohen, 1989); some of the implementation was an extension of doctoral work done at Northwestern University (Cohen, 1990) under Gary Kendall; some of the integration was done at the Human Interface Technology Lab at the University of Washington (Cohen, 1991) under Tom Furness and Woody Barfield; some of the extensions and analysis were done at the Human Interface Labs of NTT under Nobuo Koizumi (Koizumi, Cohen, & Aoki, 1992).

References

Baecker, R., Small, I., & Mander, R. (1991). Bringing icons to life. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of CHI, ACM Conference on Computer-Human Interaction,* New Orleans, LA. Reading, MA: Addison-Wesley.

Begault, D. R. (1991). Preferred sound intensity increase for sensation of half distance. *Perceptual and Motor Skills, 72,* 1019–1029.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 012

Begault, D. R., & Wenzel, E. M. (1992). Techniques and applications for binaural sound manipulation in man-machine interfaces. *International Journal of Aviation Psychology, 2*(1), 1–22.

Bernardini, N., & Otto, P. (1989). Trails: An interactive system for sound location. In *Proceedings of the International Computer Music Conference* (pp. 29–33), San Francisco, CA: Computer Music Association.

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound localization.* Cambridge, MA: The MIT Press.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustic Society of America,* **25**(5), 975–979.

Chowning, J. M. (1977). The simulation of moving sound sources. *Computer Music Journal, 1*(3), 48–52.

Cohen, M. (1987). Stereotelephonics. Internal Memorandum IM-000-21460-87-04, Bell Communications Research.

Cohen, M. (1989). Multidimensional audio window management. Technical Memorandum TM-NPL-015362, Bell Communications Research.

Cohen, M. (1990). *Multidimensional audio windows: Extending user interfaces through the use of spatial auditory information.* Ph.D. thesis, Northwestern University.

Cohen, M. (1991). Two-dimensional audio windows: Conferences, concerts, and cocktails. Technical Report HITL-R-91-1, Human Interface Technology Lab, FJ-15, University of Washington, Seattle, WA.

Cohen, M., & Koizumi, N. (1991a). Audio windows for binaural telecommunication. In *Proceedings of the joint meeting of human communication committee and speech technical committee* (pp. 21–28), Tokyo. Institute of Electronics, Information and Communication Engineers. Vol. 91, No. 242; SP91-51; HC91-23.

Cohen, M., & Koizumi, N. (1991b). Audio windows for sound field telecommunication. In *Proceedings of the seventh symposium on human interface* (pp. 703–709), Kyoto. Society of Instrument & Control Engineers, 2433.

Cohen, M., & Koizumi, N. (1992a). Iconic control for audio windows. In *Proceedings of the eighth symposium on human interface* (pp. 333–340), Kawasaki, Japan. Society of Instrument & Control Engineers. 1411.

Cohen, M., & Koizumi, N. (1992b). Exocentric control of audio imaging in binaural telecommunication. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* (Special Section on Fundamentals of Next Generation Human Interface), E75-A(2):164–170. 0916-8508.

Cohen, M., & Ludwig, L. F. (1991a). Multidimensional audio window management. *IJMMS, the Journal of Person-Computer Interaction, 34*(3), 319–336. Special Issue on Computer Supported Cooperative Work and Groupware. 0020-7373.

Cohen, M., & Ludwig, L. F. (1991b). Multidimensional audio window management. In S. Greenberg (Ed.), *Computer supported cooperative work and groupware* (pp. 193–210). London: Academic Press.

Fisher, S. (1990). Virtual environments, personal simulation & telepresence. Multidimensional Media, Aerospace Human Factors Research Division, NASA Ames Research Center, Moffett Field, CA.

Fisher, S. S., McGreevy, M., Humpries, J., & Robinett, W. (1986). Virtual environment display system. *ACM Interactive Graphics.*

Foster, S. H., Wenzel, E. M., & Taylor, R. M. (1991). Real time synthesis of complex acoustic environments. In *Proceedings of the ASSP (IEEE) Workshop on Applications of Signal Processing to Audio & Acoustics,* New Paltz, NY, summary.

Furnas, G. W. (1986). Generalized fisheye views. In *Proceedings of CHI, ACM Conference on Computer-Human Interaction* (pp. 16–23). Boston, MA: ACM.

Gehring, B. (1987). *Auditory localizer model AL-201 product description.* Gehring Research Corporation, 189 Madison Avenue, Toronto, Ontario M5R 2S6.

Gehring, B. (1988). U.S. Patent 4774515: Attitude Indicator. 189 Madison Avenue, Toronto, Ontario M5R 2S6.

Gehring, B. (1990). *Focal Point™ 3-D sound user's manual.* Gehring Research Corporation, 189 Madison Avenue, Toronto, Ontario M5R 2S6, 1.002 edition.

Johnson, J., Roberts, T. L., Verplank, W., Smith, D. C., Irby, C. H., Beard, M., & Mackey, K. (1989). The Xerox Star: A Retrospective. *(IEEE) Computer, 22*(9), 11–29.

Kendall, G. S. (1990). Visualization by ear: Auditory imagery for scientific visualization and virtual reality. In A. Wolman & M. Cohen (Eds.), *Dream machines for computer music proceedings* (pp. 41–46). Evanston, IL: School of Music, Northwestern University.

Kendall, G. S., & Freed, D. J. (1988). *Scientific visualization by ear.* Technical report, Northwestern Computer Music, Northwestern University, Evanston, IL.

Kendall, G. S., & Martens, W. L. (1984). Simulating the cues of spatial hearing in natural environments. In *Proceedings of the international computer music conference,* (pp. 111–126), Paris. Computer Music Association.

Kendall, G. S., Martens, W. L., Freed, D. J., Ludwig, M. D.,

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 013

& Karstens, R. W. (1986a). Image model reverberation from recirculating delays. In *Proceedings of the 81st Audio Engineering Society Convention.* New York: Audio Engineering Society.

Kendall, G. S., Martens, W. L., Freed, D. J., Ludwig, M. D., & Karstens, R. W. (1986b). Spatial Processing Software at Northwestern Computer Music. In *Proceedings of the International Computer Music Conference* (pp. 285–292). San Francisco, CA: Computer Music Association.

Kendall, G. S., Martens, W. L., & Wilde, M. D. (1990). A spatial sound processor for loudspeaker and headphone reproduction. In *Proceedings of the AES 8th International Conference.* New York, NY: Audio Engineering Society.

Koenig, W. (1950). Subjective effects in binaural hearing. *Journal of the Acoustic Society of America, 22*(1), 61–62.

Koizumi, N., Cohen, M., & Aoki, S. (1992). Design of virtual conferencing environments in audio telecommunication. In *Proceedings of the 92nd Audio Engineering Society Convention,* Wien, Austria. 4CA1.04, preprint 3304.

Loomis, J. M., Hebert, C., & Cininelli, J. G. (1990). Active localization of virtual sounds. *Journal of the Acoustic Society of America, 88*(4), 1757–1763.

Ludwig, L. F., & Pincever, N. C. (1989). Audio windowing and methods for its realization. Technical Memorandum TM-NPL-015361, Bell Communications Research.

Ludwig, L. F., Pincever, N. C., & Cohen, M. (1990). Extending the notion of a window system to audio. *(IEEE) Computer* (Special Issue on Voice in Computing), *23*(8), 66–72.

Martel, A. (1986). The SS-1 sound spatializer: A real-time MIDI spatialization processor. In *Proceedings of the International Computer Music Conference,* (pp. 305–307). San Francisco, CA: Computer Music Association.

Martens, W. (1989). Spatial image formation in binocular vision and binaural hearing. In *Proceedings of the 3D Media Technology Conference,* Montréal, Québec.

Martens, W. L. (1987). Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proceedings of the International Computer Music Conference.* San Francisco, CA: Computer Music Association.

McKinley, R. L., & Ericson, M. A. (1988). Digital synthesis of binaural auditory localization azimuth cues using headphones. *Journal of the Acoustic Society of America, 83,* S18.

McLuhan, H. M., & Fiore, Q. (1967). *The medium is the message.* Random House.

Metaresearch (1991). *Digital ears user manual.* Metaresearch, Inc., 516 SE Morrison, Suite M-1; Portland, OR 97214, 1.0 edition.

Moore, F. R. (1983). A general model for spatial processing of sounds. *Computer Music Journal, 7*(3), 6–15.

Myers Laboratories (1987). Information package. 20 South Santa Cruz Avenue, Suite 308, Los Gatos, CA 95030.

Richter, F., & Persterer, A. (1989). Design and applications of a creative audio processor. In *Proceedings of the 86th Convention of the AES.* Audio Engineering Society. Preprint 2782 (U-4).

Scott, D. (1989). A processor for locating stationary and moving sound sources in a simulated accoustical environment. In *Proceedings of the International Computer Music Conference* (pp. 277–280). San Francisco, CA: Computer Music Association.

Setton, M. (1990). Sonic Navigator™. Project Report, Berkeley Systems, Inc., 1700 Shattuck Ave., Berkeley, CA 94709.

Sorkin, R. D., Wightman, F. L., Kistler, D. J., & Elvers, G. C. (1989). An exploratory study of the use of movement-correlated cues in an auditory head-up display. *Human Factors, 31*(2), 161–166.

Wenzel, E. M. (1992). Localization in virtual acoustic displays. *Presence: Teleoperators and Virtual Environments, 1*(1), 80–107.

Wenzel, E. M., Stone, P. K., Fisher, S. S., & Foster, S. H. (1990). A system for three-dimensional acoustic "visualization" in a virtual environment workstation. In *Proceedings of the First IEEE Conference on Visualization* (pp. 329–337), San Francisco, CA.

Wenzel, E. M., Wightman, F. L., & Foster, S. H. (1988a). Development of a three-dimensional auditory display system. In *Proceedings of CHI, ACM Conference on Computer-Human Interaction,* Washington, DC.

Wenzel, E. M., Wightman, F. L., & Foster, S. H. (1988b). A virtual display system for conveying three-dimensional acoustic information. In *Human Factors Society 32nd Annual Meeting* (pp. 86–90). Santa Monica, CA.

Wenzel, E. M., Wightman, F. L., & Kistler, D. J. (1991). Localization of non-individualized virtual acoustic display cues. In S. P. Robertson, G. M. Olson, & J. S. Olson (Eds.), *Proceedings of CHI, ACM Conference on Computer-Human Interaction,* New Orleans, LA. Reading, MA: Addison-Wesley.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/4/468/1622533/pres.1992.1.4.468.pdf by Reprints Desk on 01 December 2022

Meta Platforms, Inc.
Exhibit 1052
Page 014